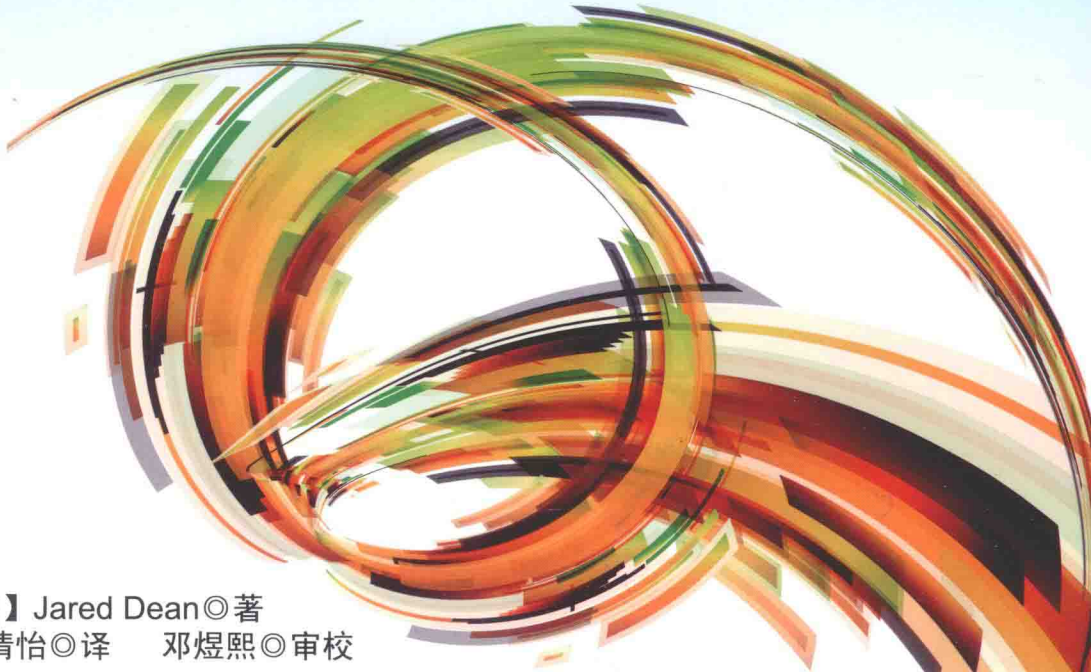
“十二五”
国家重点图书出版规划项目

sas WILEY

新信息时代商业经济与管理译丛



【美】Jared Dean◎著
林清怡◎译　邓煜熙◎审校

大数据挖掘与机器学习

工业4.0时代重塑商业价值

BIG DATA, DATA MINING AND MACHINE LEARNING

VALUE CREATION FOR BUSINESS LEADERS AND PRACTITIONERS

中国工信出版集团　人民邮电出版社 POSTS & TELECOM PRESS

“十二五”
国家重点图书出版规划项目

新信息时代商业经济与管理译丛

大数据挖掘与机器学习

工业4.0时代重塑商业价值

【美】Jared Dean◎著
林清怡◎译　邓煜熙◎审校

BIG DATA, DATA MINING AND MACHINE LEARNING

VALUE CREATION FOR BUSINESS LEADERS AND PRACTITIONERS

人 民 邮 电 出 版 社
北　京

图书在版编目（CIP）数据

大数据挖掘与机器学习 ：工业4.0时代重塑商业价值/（美）迪安（Dean, J.）著 ；林清怡译. -- 北京 ：人民邮电出版社，2015.10
（新信息时代商业经济与管理译丛）
ISBN 978-7-115-39736-2

Ⅰ. ①大… Ⅱ. ①迪… ②林… Ⅲ. ①商业信息－数据处理 Ⅳ. ①F713.51

中国版本图书馆CIP数据核字(2015)第190515号

版权声明

◆ 著　　[美] Jared Dean
译　　林清怡
审　　校　邓煜熙
责任编辑　刘　洋
责任印制　彭志环
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷
◆ 开本：700×1000　1/16
印张：16　　2015 年 10 月第 1 版
字数：228 千字　　2015 年 10 月河北第 1 次印刷
著作权合同登记号　图字：01-2014-5084 号

定价：55.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京崇工商广字第 0021 号

内容提要

本书分为3个部分，共17章。第Ⅰ部分“计算环境”，包括第1章到第3章。第Ⅱ部分“将数据转化为商业价值”，包括第4章到第10章。这一部分聚焦于数据挖掘活动中所要用到的方法、算法和路径。第Ⅲ部分“将其全部结合起来的成功案例”包括第11章到第17章。本部分主要描述了作者参与过的成功应用大数据分析优化企业决策、提高企业价值的公司案例。

本书可作为企业管理人员、营销主管、分析人员、IT人员等作为理解大数据、应用大数据为企业创造价值的指引，同时，本书也可供统计学、应用数学及计算机专业学者和研究人员参考学习。

献给我的妻子，没有她的帮助和奉献，本书就不可能面世。谢谢你，Katie！

给 Geoffrey、Ava、Mason 和 Chase 的话：要记住，只有努力才能让你最快到达舒适之地。

专家赞誉

“Jared 的书是对高性能分析领域的伟大指引。对于那些具有预测分析经验但需要对如何应用技术提升已有方法并且创造出新的可能性方面更加娴熟的人来说，这本书非常有用。对于需要为下一代高级分析构建环境并从中获益的业务高管和 IT 专业人员，也能从中受益。”

——Jonathan Levine，万豪国际（Marriott International）
消费者洞察分析资深总监

“Jared 描述的观点与我们 Kaggle 竞赛获奖者所应用的观点相同。对于那些想要更深入学习并充分理解数据挖掘、知识探索和从数据中提取价值的方方面面的人来说，本书是一个很好的综述。”

——Anthony Goldbollm，Kaggle 创始人兼 CEO

“Jared 在本书中提出的概念对于我教的那些学生极其有价值，可以帮助他们更充分地理解当组织开始利用其数据时可以开启的能力。书中的案例研究特别有用，有助于学生们获得能够做什么的愿景。Jared 对分析的热情在其字里行间显现，他完成了使复杂观点对层次各异的读者更平易近人的伟大工作。”

——Tonya Etchison Balan，博士，北卡罗来纳州立大学
商学院统计学教授

致谢

我要感谢所有帮助我让这本书面世的那些人。这是一段漫长的旅程，也是一个极好的学习和成长的体验过程。

Patrick Hall，谢谢你对我观点的认可以及来自你个人的诸多贡献。我很感激你能够跟我讨论观点和趋势，并且提供深思熟虑的、及时的、有用的反馈。

Joseph Pingenot、Ilknur Kabul、Jorge Silva、Larry Lewis、Susan Haller 和 Wendy Czika，感谢你们与我分享行业知识和对分析的热爱。

Michael Wallis，谢谢你在文本分析领域和“危险边缘”案例开发中提供的帮助。

Udo Sglavo 和 Taiyeong Lee，谢谢你们对时间序列数据挖掘分析的评论以及所提供的巨大帮助。

Barbara Walers 和 Vicki Jones，谢谢你们阅读硬件如何影响软件这一部分，并反馈你们的理解。

Jared Peterson，感谢你帮助我从 Nike+Fuelband 智能手环中下载数据。

Franklin So，感谢你对客户核心业务问题的精彩描述。

谢谢 Catherine Coyne 奶奶，您牺牲了许多时间帮助一位年轻作者编辑稿件，大大提高了稿件的可读性。我对您的帮助充满感激，希望在我 80 岁的时候，我能够有您一半的活力。

我还要感谢 SAS 公司和 John Wiley&Sons 出版社的人员，在这个项目的各个阶段，包括在此过程里出现的几个主要弯路中，你们都给予了我反馈和支持。

最后，我还要感谢我的妻子 Katie，在我研究、撰写、编校、继续撰写的过程中担负了许多责任。遇到你是我一生中最美好的事情。

序言

我喜欢分析预测这个领域，我的整个职业生涯都在这个世界中。数学非常有趣（至少对我而言），但是将算法发现的东西转化为解决方案，为公司所用并令其能从中获利，会使数学更有价值。从某些方面来看，Jared Dean 和我在这点上有点不同寻常：我们确实非常喜欢看到这些解决方案能对我们为之工作的组织起作用。但是令我们惊讶的是，我们过去经常坐在办公室后面埋头苦干的这个领域，现在已经逐渐变成 21 世纪最为性感的工作之一。这一切是怎么发生的呢？

在我们生活的世界中，数据采集的规模日益增长，能够对人们和机器的主要所为进行归纳总结，并以更细微的粒度捕捉其行为。描述数据特征的这 3 种方式有时候被称作规模、种类和速度——这也是对大数据的界定方式。它们之所以被采集是因为人们感觉到在数据中存在着价值，即使人们并不确切知道将用这些数据做什么。最初，许多组织采集数据，对其总结、报告，通常是应用来自于商业智能的方法，而这在现在已经司空见惯。

但是在近年来，出现了模式上的转变。组织已经发现分析预测改变了他们的决策方式。本书中描述的预测建模的算法和方法大多都不是崭新的，Jared 自己也将大数据问题描述为没啥新意。他所描述的算法全都至少已经有 15 年之久了，对其有效性的见证表明根本就不需要新算法。虽然如此，当许多公司试图用数据优化决策时，预测建模确实令其耳目一新。

这些公司不仅需要理解预测建模的技术和原理，还需要理解如何在对标准方法和回答形成挑战的问题上应用这些原则。

但是对于预测建模，除了构建预测模型，还有更多的工作。预测建模项目的运作方面经常被忽略，在书本和课程中都很少被覆盖。首先，这包括指定预测建模所需要的硬件和软件。就如 Jared 所描述的那样，这要由组织、数据和这个项目的分析师所决定。如果未能提供给分析师适当的资源，项目就会步履蹒跚，通常归于失败。我本人就曾在我工作的项目中遇到过这样的情况，因为指定的硬件不恰当导致我花了相当多时间来解决在 RAM 和处理速度上的局限。

最终，预测建模项目的成功是通过对于应用它的组织至关重要的那些指标来进行衡量的，即它是否提高了效率、ROI、客户生命期价值，或者还包括对一些软指标如公司信誉的衡量。我很喜欢这本书中关于处理这些问题的案例研究，这里有半打案例可以让你胃口大开。对于试图理解预测建模如何影响其基线的经理人员而言，这特别重要。

预测建模是科学，但是对预测建模解决方案的成功部署需要将模型与业务相关联。对于认识这些关联，经验是至关重要的，你可以从这里提取经验财富，驱使你在预测建模的旅程中继续向前。

Dean Abbott

Abbott Analytics，Inc.

2014 年 3 月

中文版序

在移动互联网时代，社交网络成为推动移动互联网迅猛发展的生力军。互联网花了 30 年时间达到 7.5 亿用户，成立于 2004 年的 Facebook 只花了 8 年时间便达到与之不相上下的用户数。

社交网络的核心价值在于人和人的社交关系，马克·扎克伯格说："人们分享得越多，他们就能够通过自己信赖的人，获得更多有关产品和服务的信息。他们能够更加轻松地找到最佳产品，并提高生活品质和效率。在这一过程中，企业获得的益处是，他们能够制造更好的产品，即以人为本的个性化产品。与传统商品相比，那些基于社交关系、社交图谱、社交圈推广的产品更富有吸引力。"可见，社交网络为人们开拓了新的信息分享和交流空间，也为企业创造了利用社交关系更开阔、更深入、更高效地开展客户销售、服务和营销的机会。对于企业来说，谁更早抓住机会研究了解自身的客户社交网络关系，谁就更具核心市场竞争力。

博雅公关 Burson-Marsteller 和互联网监测分析公司 Visible 联合发布的 2012 年度财富 100 强公司社会化媒体使用报告显示，2010～2012 年，100 家公司平均拥有 Twitter 账号分别为 4.2 个、5.8 个和 10.1 个，Facebook 账号分别为 2.1 个、4.2 个和 10.4 个，YouTube 账号分别为 1.6 个、2.7 个和 8.1 个。而根据 LinkedIn 与市场研究公司 TNS 于 2014 年 2 月发布的合作研究成果，在美国中小型企业中，81%的被调查者使用社交媒体促进业务增长，94% 将社交媒体作为营销工具，而 49% 为了教育目的使用社

交媒体，并获取业务洞察力。可见，确实如制订企业社会化媒体实践“黄金标准”、著有《营销和公共关系的新规则》一书的营销专家大卫•米尔曼•斯科特（David Meerman Scott）所言：“我们正在经历一场人类沟通方式的变革。我认为这是自印刷机发明以来人类沟通方式最显著的革命……社会化媒体已经在革命性地改变商业沟通。”

我们知道，这是移动互联网时代，这是社交网络时代，而同时，人们的数字化生存让有关人们生活甚至工作的行为信息都数字化，而这些以单个个体为对象的形形色色、包罗万象、细致入微、支撑个体兴趣需求和喜好的数字化信息构成了大数据，所以，这个时代更是一个大数据时代：到今天，世界上所有印刷材料的数据量是 200PB，全人类说过所有对话的数据量大约是 5EB；每天我们产生的数据大约是 2.5PB，这就意味着当今世界全部数据的 90%都在近两年产生。

如果我们有相应的 IT 技术、分析手段驾驭大数据，大数据就是金矿；如果没有相应的技术和手段，大数据就将成为淹没我们的海洋。谈论大数据在整个社会确实已成为一种时髦，但是根据麦肯锡在 2012 年 4 月的调查，仅有 1/5 的受访者所在公司已经在一个业务单元或职能部门完全部署大数据和分析，以获得客户洞察；仅有 13% 的受访者表示，公司全面使用数据获得洞见。可见，大数据要从谈论和研究到技术和应用实现，路途还很漫长，所以，如何客观审慎地对待已有的大数据优势，提前思考并规划、架构、完善、部署数据从采集、清洗、存储、分析、应用以及管理监控的全企业层面的 BI（商业智能）平台，并培养贯穿企业运营管理流程的 BA（商业分析）体系，用数据说话，实现全企业层面的精确管理和精确营销、销售、服务，也就是大数据时代我们最终能够成为时代弄潮儿抑或被潮水淹没者的“To be or not to be”的关键问题。

中国电信股份有限公司广州研究院市场运营研究所，长年从事电信企业运营管理及市场研究的实践和方法总结，研究时间最长的已达 17 年，并分别在行业竞争、商业模式创新、精确营销、品牌、舆情、口碑营销、数据分析及挖掘、数据仓库/BI 架构及规范等细分领域长年支撑企业运营管理实践，不仅对企业运营有深刻理解和独到见解，且基于企业运营

管理实践完成了大量方法创新和应用研究，发表了多本论著和数百篇专业论文，为各细分专业领域积累了众多的方法、经验和模型。

近几年，随着移动互联网—社交网络—大数据的迅猛发展，也因为企业转型的需要，市场运营研究所在邓煜熙所长带领下，研究人员围绕两大问题开展相关研究：（1）企业如何建立自己的社交媒体策略并进行社交网络分析；（2）为实现精确管理、精确营销、销售和服务，企业如何架构 BI 平台和 BA 体系。部门集中有关资源，有计划、有步骤、层层推进地深入开展研究，完成相关科研项目和撰写论文若干。

接下来，围绕客户关系管理、客户体验管理大体系，以支撑企业生产运营管理流程各环节运作，我们预计对企业大数据体系架构和分析、应用等方面进行深入研究。

最后，借狄更斯的话，“这是最好的时代，也是最坏的时代；这是智慧的年代，也是愚蠢的年代；这是信仰的时期，也是怀疑的时期；这是光明的季节，也是黑暗的季节；这是希望的春天，也是失望的冬天；大伙儿面前应有尽有，大伙儿面前一无所有”，让大伙儿一起，掌握商业智能、商业分析两大工具，驾驭社交媒体，洞察社交网络，弄潮大数据。

中国电信股份有限公司广州研究院院长　蔡康

2015 年 7 月于广州

前言

我在 SAS 公司担任现在这个数据挖掘开发管理角色的第一周时，首次接到撰写本书的项目。写一本书总会列于死前必做的清单中，我非常兴奋于能参与其中。我开始认识到为何希望写书的人这么多，但有机会看到他们的想法和观点能装订出版的人却这么少。

在我的学习和职业生涯中，我很幸运能有机会处于数据挖掘领域一些伟大进展的前沿和中心，也有机会在某些杰出人士的指导下学习研究。这一经历能帮助我定位在创建这方面工作时所需要的技能和经验。

数据挖掘是一个我热爱的领域。从孩提时代开始，我就渴望能解释各种事物的工作原理，渴望理解系统在“普通”情况和极端情况下是如何运作的。从小学到中学，我一直认为工程师将会是能将我的好奇心和我对于解释周边世界的渴望相结合的工作。但是，在我本科生的最后一年，我发现了统计和信息系统，并为之着迷。

在本书第一部分，我探索了硬件基础和系统架构。我父母非常慷慨，允许我沉溺于这个我喜爱的领域中，在当时，购买计算机的费用可远不止 299 美元。我家里的第一台计算机是 Apple IIc，有两个 5.25 英寸软驱，没有硬盘驱动器。几年后，我用工具包组装了一台 Intel 386 PC，我对于玩电脑游戏以及点击加速键将 CPU 主频速度从 8MHz 提高到 16MHz 记忆犹新。我亲眼目睹了摩尔定律，但是仍然为智能手机比在水星计划、

阿波罗空间计划和轨道航天飞行计划中所用到的计算机加起来的计算能力更强而感到吃惊。

在我获得统计学学士学位之后，我开始在美国人口统计局为联邦政府工作。这是我第一次接触到大数据的地方。在加入统计局之前，我还从来没有写过运行时间超过 1 分钟的计算机程序（除非是要求程序运行超过 1 分钟）。我的最初项目之一是处理主地址文件（MAF，Master Address File）[1]，这是一个由统计局维护的地址列表。这个地址列表也是人口统计局执行近期调研的主要调查框（是的，另外 9 年里当然还有许多其他工作）。这个列表有超过 3 亿条记录，集成了所有地址信息、经度信息和地理信息，有数百个与每个住宅单元相关联的属性。在处理这样一个大型数据集时，我第一次认识到编程效率、可扩展性和硬件优化。我非常感激我富有耐心的经理 Maryann，她给了我学习的时间，交给我有趣而又宝贵的项目，这些项目给了我实践的经验和创新机会。这是一个很棒的位置，因为我接触到了以往在学院中从未学习过的新的技术和方法。对于任何新项目，总会有某些观点效果很好而其他观点则流于失败。我参与的一个具体项目是试图识别出在 2000 年人口调查中哪些街区（人口统计局将美国划分成独立的地理区域——其层级是州、县、道、片区、街区。在美国大概有 820 万个街区）被高估、哪些被低估。从已有数据中，我们无法找到一个方法，能够证明我们用于预测实际住宅单元数量与报告住宅单元数量之间偏差的模型是精确的。这个项目非常幸运，它从国会获得基金，能够进行实地研究，提供对模型的反馈和检验。这是我第一次听到“数据挖掘”这个术语，也是第一次接触到 SAS™ Enterprise Miner®和 Salford 系统的 CART®。经过在人口统计局一段时间的工作，我意识到我需要接受进一步的教育才能实现我的职业目标，所以我进入弗吉尼亚州菲尔菲克斯（Fairfax）县乔治梅森大学（George Mason University）统计系学习。

在研究生院，我更加详细地学习了在数据挖掘、机器学习和统计领域的常见算法，包括生存分析、调研抽样和计算统计学。在我的研究生学习中，我能够将这些课堂上学到的课程结合到办公室所需要的实际数据分

[1] MAF 是在 10 年一次的人口普查操作中为美国的每个住宅单元或潜在住宅单元创建的。

析和创新中。我获得了对不同理论及用于数据分析和预测分析的不同方法的相对优缺点的理解。

从研究生院毕业以后，我改变了自己的职业方向，从数据分析[2]角色转而成为一名软件开发人员。我为 SAS 研究院工作，在那里我参与了之前使用过的软件的设计。我从使用软件转向创建软件。我认识到 SAS 加诸于软件之上的严格的数值验证，了解到 SAS 全面的文档，意识到公司为了使得对软件的新改进能够与原有软件兼容、并且不断开发出客户需要的新的软件特性而付出的不懈努力，它们同样也给我带来了新的挑战和成长机会。

在 SAS 工作的这些年间，我逐渐能充分理解软件的设计开发方式和我们的客户对软件的应用方式。我经常有机会拜访客户，聆听他们遇到的商业挑战，并推荐能够帮助引领他们成功的方法或流程，为他们的组织创造价值。

我写的这本书就是源于我收集的这些经验，在此过程中还得到了来自 SAS 研究院内外部的出色的工作人员以及我的同事们的帮助。

---

[2] 我是一名数据科学家，早在这个术语被创造出来之前就是如此。

“新信息时代商业经济与管理译丛”简介

“新信息时代商业经济与管理译丛”是人民邮电出版社精心策划的一套重点图书，甄选世界各大著名出版商的优质外版资源，已被国家新闻出版广电总局列入“十二五”国家重点图书出版规划项目。本套译丛主要围绕“大数据”这个核心进行内容选择和扩展，包括商业智能、商业分析、数据挖掘、数据可视化、大数据营销、大数据商业应用等热点主题，以及商业模式、供应链、物流管理、精益思想等其他经管类主题。我们真诚希望本套译丛可以成为大数据时代各位读者在工作和学习中的好帮手。

如有意推荐、翻译优质图书，或有其他沟通、合作意向，都欢迎与我们联系。本丛书总策划人：刘洋，电子邮箱：liuyang@ptpress.com.cn。

精品图书

《大数据决策：商业分析新常态》

印孚瑟斯（Infosys）公司高级首席、得克萨斯A&M大学教授毕生经验集大成之作，直击大数据决策的核心实质，多位顶级咨询和营销专家联袂推荐。

定价：59元

ISBN：978-7-115-39180-3

《大数据预测：需求驱动与供应链变革》

SAS公司全球顶级专家毕生研究成果，麻省理工学院（MIT）研究员、乐高（LEGO）集团预测总监、雀巢（Nestle）需求规划高级经理、中国电信资深专家倾力推荐。

定价：69元

ISBN：978-7-115-38880-3

《大数据营销：互联网+时代如何定位客户》

美国联邦政府和IBM公司高级顾问Jeff Tanner博士潜心钻研多年之成果，Babson商学院教授、畅销书作者Teradata首席营销官等多位专家鼎力推荐！

定价：49元

ISBN：978-7-115-39241-1

《大数据可视化：重构智慧社会》

通过数据可视化，领略大数据的美丽与魅力，卡内基梅隆与康奈尔大学出身的顶级专家告诉你，身处互联网+时代，如何才能让自己所在的组织更加智能。

定价：59元

ISBN：978-7-115-39269-5